Metagenomic gene annotation by a homology-independent approach

Changbin Du[1,2], Jeff Froula[1,2], Tao Zhang[1,2], Annette Salmeen[1,2], Matthias Hess[3], Cheryl A. Kerfeld[1,2,4], Zhong Wang[1,2]

[1]Lawrence Berkeley Lab, Genomics Division, Berkeley, CA USA
[2]Department of Entergy, Joint Genome Institute, Walnut Creek, CA USA
[3]Washington State University, Department of Molecular Biosciences, Richland, WA USA
[4]Department of Plant and Microbial Biology, UC Berkeley, Berkeley, CA USA

To whom correspondence may be addressed. E-mail: cdu@lbl.gov.

March 22, 2011

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Metagenomic gene annotation by a homology-independent approach

Changbin Du[1,2], Jeff Froula[1,2], Tao Zhang[1,2], Annette Salmeen[1,2], Matthias Hess[3], Cheryl A. Kerfeld[1,2,4], Zhong Wang[1,2]

1. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
2. Department of Energy, Joint Genome Institute, Walnut Creek, California, USA
3. Washington State University, Department of Molecular Biosciences, Richland, WA, USA
4. Department of Plant and Microbial Biology, UC Berkeley, Berkeley, California, USA

I. Abstract

Fully understanding the genetic potential of a microbial community requires functional annotation of all the genes it encodes. The recently developed deep metagenome sequencing approach has enabled rapid identification of millions of genes from a complex microbial community without cultivation. Current homology-based gene annotation fails to detect distantly-related or structural homologs. Furthermore, homology searches with millions of genes are very computational intensive.

To overcome these limitations, we developed rhModeller, a homology-independent software pipeline to efficiently annotate genes from metagenomic sequencing projects. Using cellulases and carbonic anhydrases as two independent test cases, we demonstrated that rhModeller is much faster than HMMER but with comparable accuracy, at 94.5% and 99.9% accuracy, respectively. More importantly, rhModeller has the ability to detect novel proteins that do not share significant homology to any known protein families.

As ~50% of the 2 million genes derived from the cow rumen metagenome failed to be annotated based on sequence homology, we tested whether rhModeller could be used to annotate these genes. Preliminary results suggest that rhModeller is robust in the presence of missense and frameshift mutations, two common errors in metagenomic genes. Applying the pipeline to the cow rumen genes identified 4,990 novel cellulases candidates and 8,196 novel carbonic anhydrase candidates.

In summary, we expect rhModeller to dramatically increase the speed and quality of metagnomic gene annotation.

II. Motivation

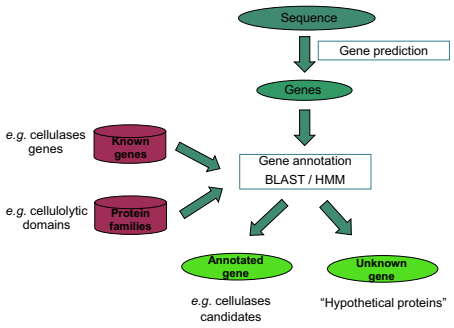

Figure 1. What homology-based gene annotation miss?

- Can not annotate genes lacking functionally characterized homologues
- Can not annotate genes with only structural similarity
- Can not identify novel or new enzymes
- The annotation may not be correct if the best match is not a true ortholog

Goal: Develop a high through-put gene annotation method for metagenomic data

III. Methods

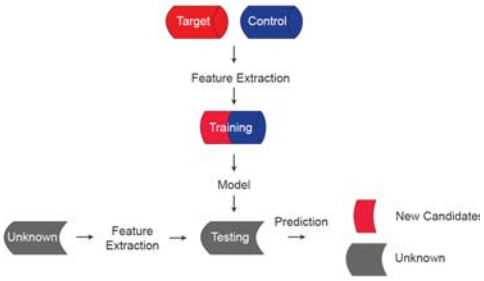

Figure 2. An overview of the workflow supported by the current version of rhModeller.

IV. Results

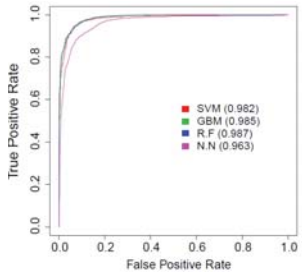

Figure 3. The features are sufficient for making predictions as shown by different algorithms.

Table 1. The rhModeller made high prediction-accuracy in two different data sets..

	Cellulases	Carbonic Anhydrases
Accuracy	0.945	0.990
False Positive Rate	0.040	0.004
False Negative Rate	0.080	0.019

Note: Values were calculated by 10-fold cross validation.

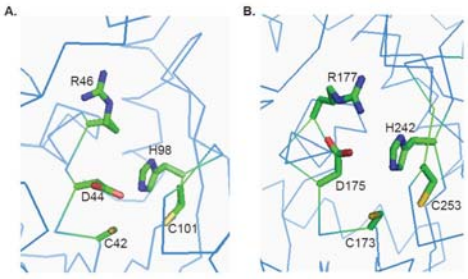

Figure 4. The rhModeller identified novel enzymes by using information of amino acid residues in active sites. A. Known β-Carbonix anhydrase. B. Novel Carbonic Anhydrase

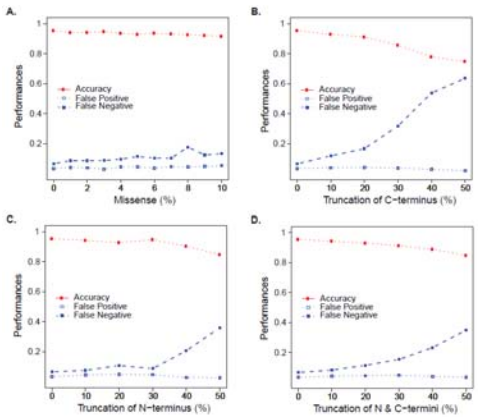

Figure 5. The rhModeller is robust against common noise in metagenomic data set.

Table 2. The rhModeller detected novel enzyme candidates from cow rumen metagenomic data

Classification	Cellulases	CAs
Target enzymes	4990 (0.01)	8196 (0.02)
Controls	499052 (0.99)	495846 (0.98)
Total	504042	504042

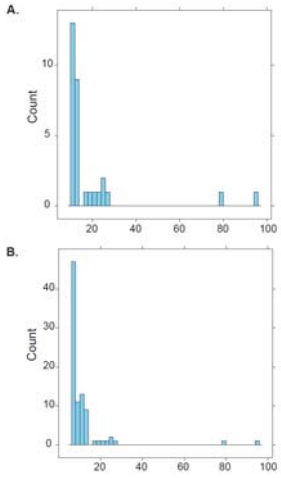

Figure 7. The rhModeller detected novel enzymes families from cow rumen metagenomic data. A. Cluster size larger than 10. B. Cluster size larger than 5.

V. Conclusions

- Novel enzymes with remote homology can be identified by rhModeller.
- The rhModeller made high prediction-accuracy on testing cases.
- The pipeline can generalize to discover new enzymes of other families.
- Our pipeline will increase the speed and quality of metagenomic gene annotation.

VI. Acknowledgements

Mingkun Li, Dongying Wu, Alexander Sczyrba, Iain Anderson & IMG

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.